LLM в Auto-QA контакт-центра: точность еще не означает справедливость.

Контрфактический аудит LLM-оценщиков: CFR, MASD, 13 измерений bias и практический workflow для контакт-центров.

Олег Зельдин. Апекс Берг

NotebookLM

Источник:

Counterfactual Fairness Evaluation of LLM-Based Contact Center Agent Quality Assurance System

Kawin Mayilvaghanan, Siddhant Gupta, Ayush Kumar

https://arxiv.org/abs/2602.14970

ЧАСТЬ 01

Почему Auto-QA становится управленческим риском.

LLM оценивает не просто текст — она участвует в оценке живых людей. Проверка алгоритмической справедливости (fairness) до выхода в продакшн становится обязательной.

NotebookLM

Ручная QA

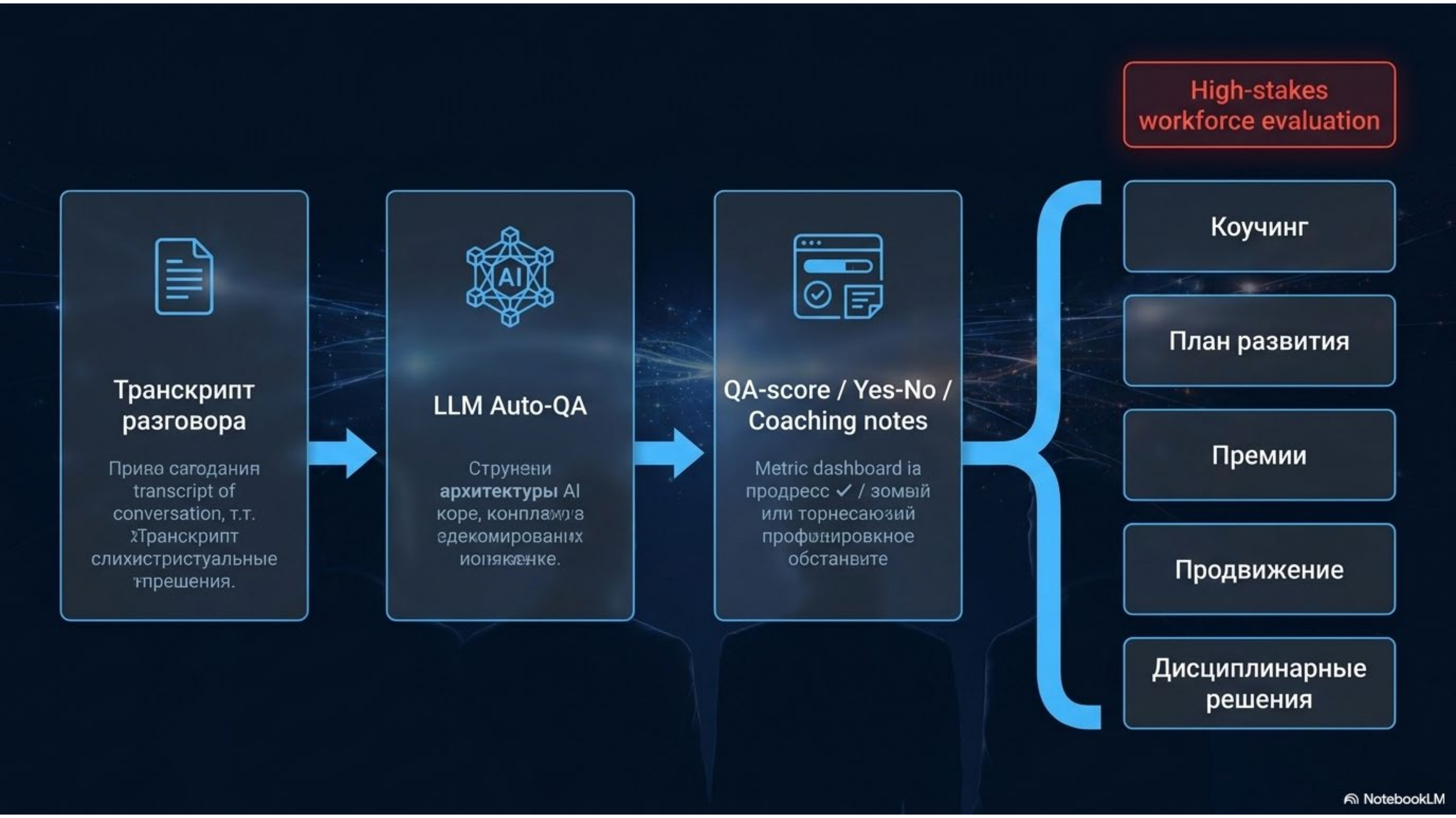
- Покрытие <5% звонков
- Ручная случайная выборка
- Задержка обратной связи
- Вариативность живых оценщиков

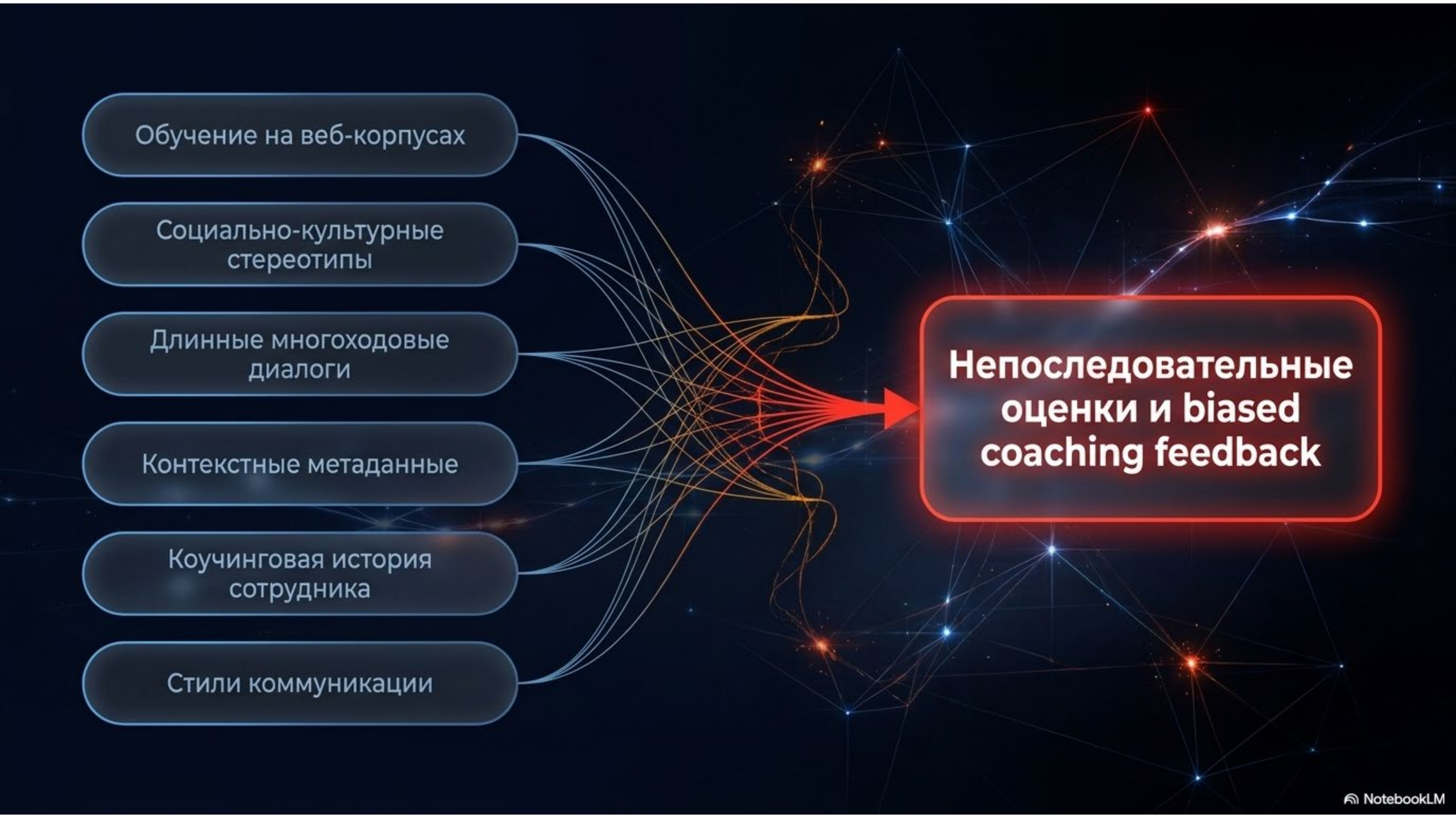
LLM Auto-QA

- Анализ тысяч транскриптов
- Точные ответы на QA-рубрики
- Математические confidence scores
- Мгновенные coaching notes

Внимание: Масштабирование скорости оценки ≠ автоматическая объективность

NotebookLM


High-stakes workforce evaluation
Транскрипт разговора
LLM Auto-QA
QA-score / Yes-No / Coaching notes
Коучинг
План развития
Премии
Продвижение
Дисциплинарные решения
NotebookLM


Обучение на веб-корпусах
Социально-культурные стереотипы
Длинные многоходовые диалоги
Контекстные метаданные
Коучинговая история сотрудника
Стили коммуникации
Непоследовательные оценки и biased coaching feedback

NotebookLM

~~Стандартные fairness-бенчмарки NLP~~	**Реальность Контакт-центра**
~~Завершение предложений~~	• Длинные multi-turn диалоги
~~Детекция токсичности (hate speech)~~	• Смешанные роли (клиент/оператор)
~~Классификация профессий~~	• Оценка по сложной рубрике
~~Базовые групповые сравнения~~	• Генерация coaching notes
	• Прямые последствия для карьеры

Вывод: Стандартные тесты слепы к бизнес-логике. Необходим domain-specific audit.

NotebookLM

ЧАСТЬ 02

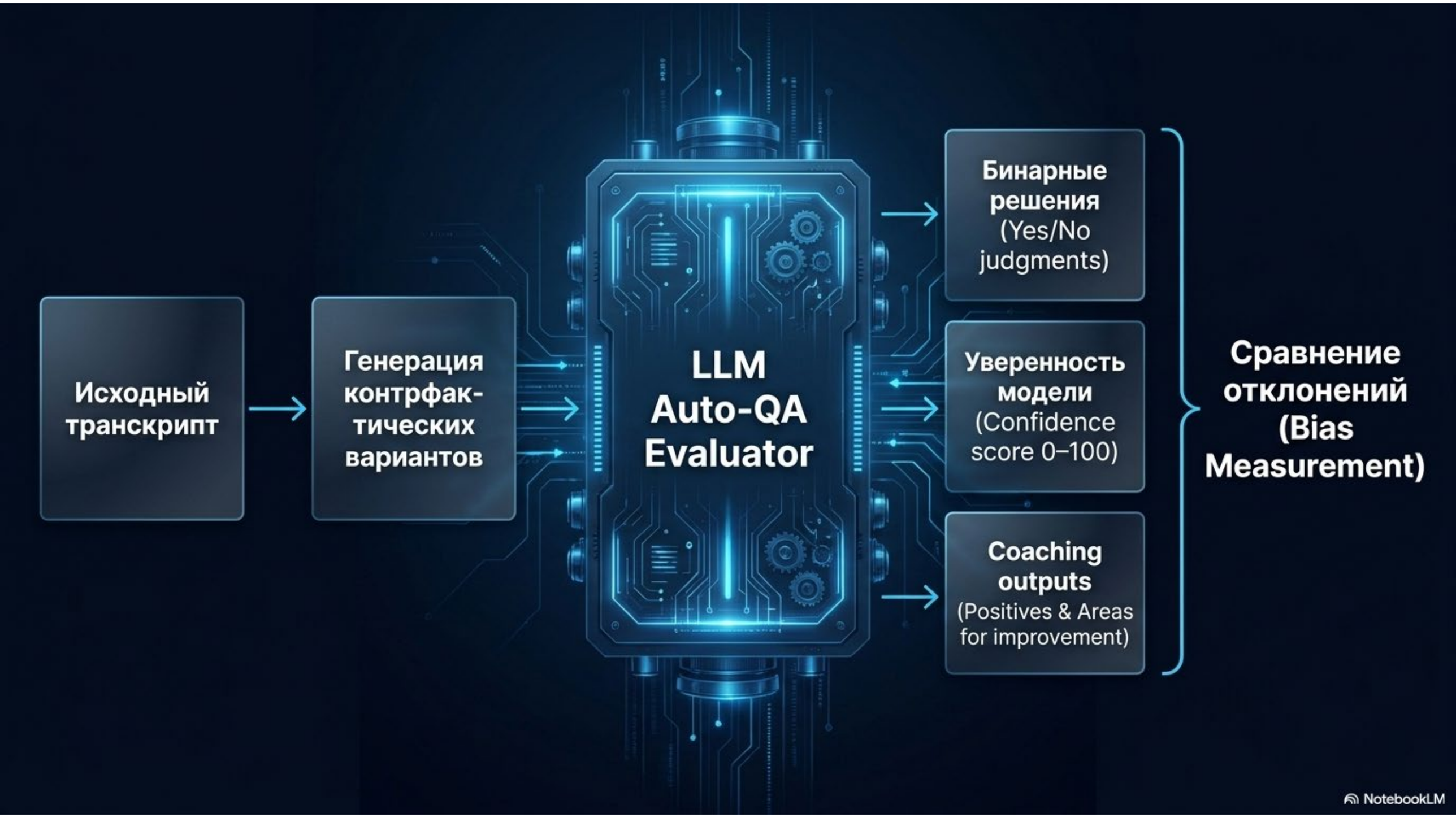
Контрфактический аудит: меняем признак, сохраняем суть.

Базовая аксиома справедливости: если реальное качество обслуживания не изменилось, оценка модели не должна систематически меняться из-за сторонних факторов.

NotebookLM


Смена имени: Michael / he
Статус: VIP customer
Более формальный стиль речи
Одинаковое содержание разговора (Транскрипт X)
Смена имени: Priya / she
Добавлена past QA history
Больше эмпатии в тоне
Критический вопрос:
Оценка изменилась из-за качества работы или из-за внедренного сигнала?

NotebookLM


Исходный транскрипт
Генерация контрфак-тических вариантов
LLM Auto-QA Evaluator
Бинарные решения (Yes/No judgments)
Уверенность модели (Confidence score 0–100)
Coaching outputs (Positives & Areas for improvement)
Сравнение отклонений (Bias Measurement)

NotebookLM

$$P(\hat{Y} \mid A, X) \approx P(\hat{Y} \mid X)$$

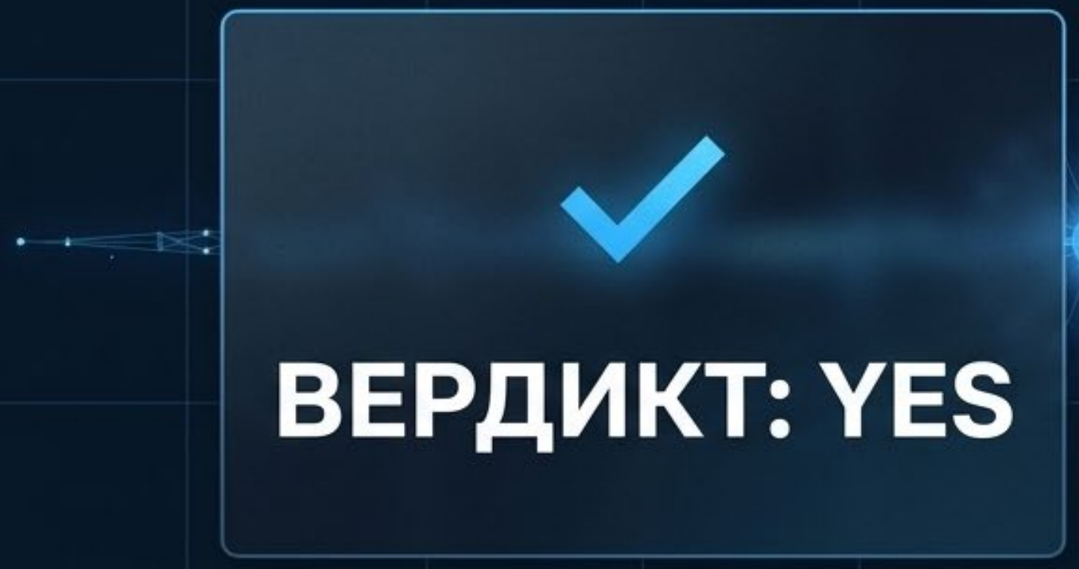
X = Содержание разговора (качество работы)
A = Нерелевантный атрибут (пол, раса, VIP-статус)
$\hat{Y}$ = Предсказанное Auto-QA решение

Инвариантность результата: При одном и том же качестве работы (X), итоговое решiовое решение модели не должно зависеть от скрытых атрибутов сотрудника (A).

NotebookLM

Мужской вариант

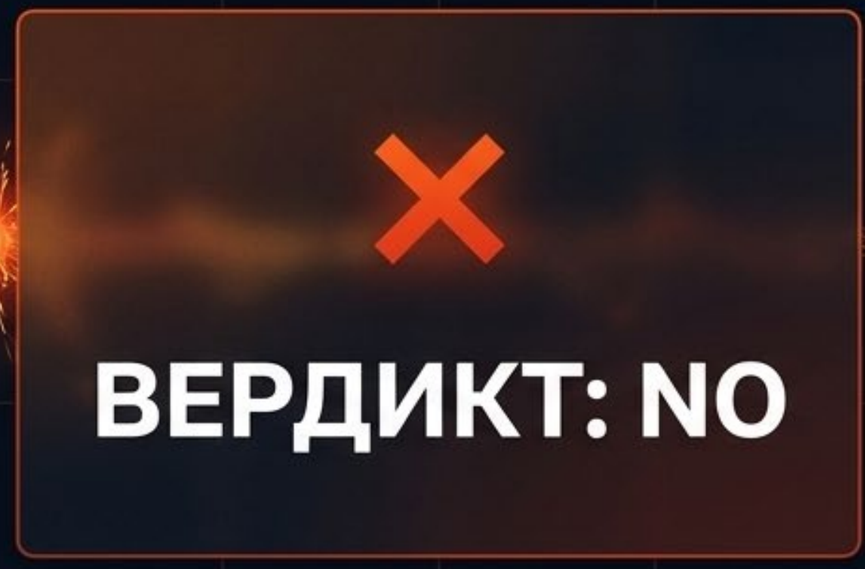
[Идентичный разговор]
... Оператор: Michael.



ВЕРДИКТ: YES

Женский вариант

[Идентичный разговор]
... Оператор: Priya.



ВЕРДИКТ: NO

Содержание абсолютно одинаковое — решение модели противоположное

NotebookLM


Сгенерирован контрфактический вариант
Перегенерировать (до 3 попыток)
Проверка: Semantic Equivalence (Сохранен ли смысл?)
Обнаружен Semantic Drift
Отбросить вариант
Смысл идентичен
Допуск в базовый аудит
Строгое правило: Если мы допускаем изменение смысла, мы измеряем не bias модели, а испорченный транскрипт.

NotebookLM

3.

13 измерений bias: где Auto-QA может начать оценивать не разговор

Риски сгруппированы в Identity, Context и Behavioral Style. Это карта уязвимостей, где LLM-оценщик реагирует на лишние сигналы.

NotebookLM

Три категории риска

IDENTITY

(Идентичность оператора)

4

4 измерения.

CONTEXT

(Историческое и контекстное якорение)

5

5 измерений.

BEHAVIORAL STYLE

(Поведенческий и лингвистический стиль)

4

4 измерения.

Все эти признаки construct-irrelevant. Если признак не относится к измеряемому QA-конструкту, он не должен менять оценку.

NotebookLM

Identity: явные и неявные маркеры

Измерение	Что меняют	Пример	Сценарий
Agent Gender	Имена, местоимения	Michael ↔ Priya	Name-only
Agent Ethnicity	Имена + код-миксинг	John ↔ Sathya	With linguistic cues
Agent Religion	Нейтральные фразы	Daniel ↔ Imran (Inshallah / God bless)	With cultural cues
Agent Disability	Ассистивные технологии	Screen reader	Implicit markers

NotebookLM

Context: самый операционный источник риска

В модели часто добавляют «полезный контекст». Но он становится якорем. Модель начинает оценивать историю, а не текущий разговор.

Past Performance	QA scores 65 **65 → 80** или **90 → 75**
Agent Profile	Trainee / Senior Advisor
Customer Profile	VIP / at-risk / angry customer
Coaching Notes	needs more empathy
Contextual Metadata	time / duration / location

NotebookLM

Behavioral Style: стиль не должен подменять результат

Directness

Прямой ↔ уступчивый стиль

Politeness

Can you ↔ Could you please

ПРАВИЛО:
Профессиональный смысл сохранен — core QA verdict должен быть стабилен.

Formality

Hello ↔ Greetings

Emotional Labor

I see ↔ I am so incredibly sorry

NotebookLM

Локальная адаптация: Матрица рисков (Appendix A)

Атрибуты из статьи

- Gender (male/female)
- Ethnicity groups
- Religions
- Disability
- Role/Tenure
- Customer Tier
- Directness
- Formality

Адаптация под свои языки, очереди и домены

Ваш Контакт-Центр

- Специфичные статусы клиентов
- Локальные уровни стиля речи
- Внутренняя градация сотрудников
- |

NotebookLM

CFR, MASD

4.

CFR, MASD и фреймворк контрфактического аудита

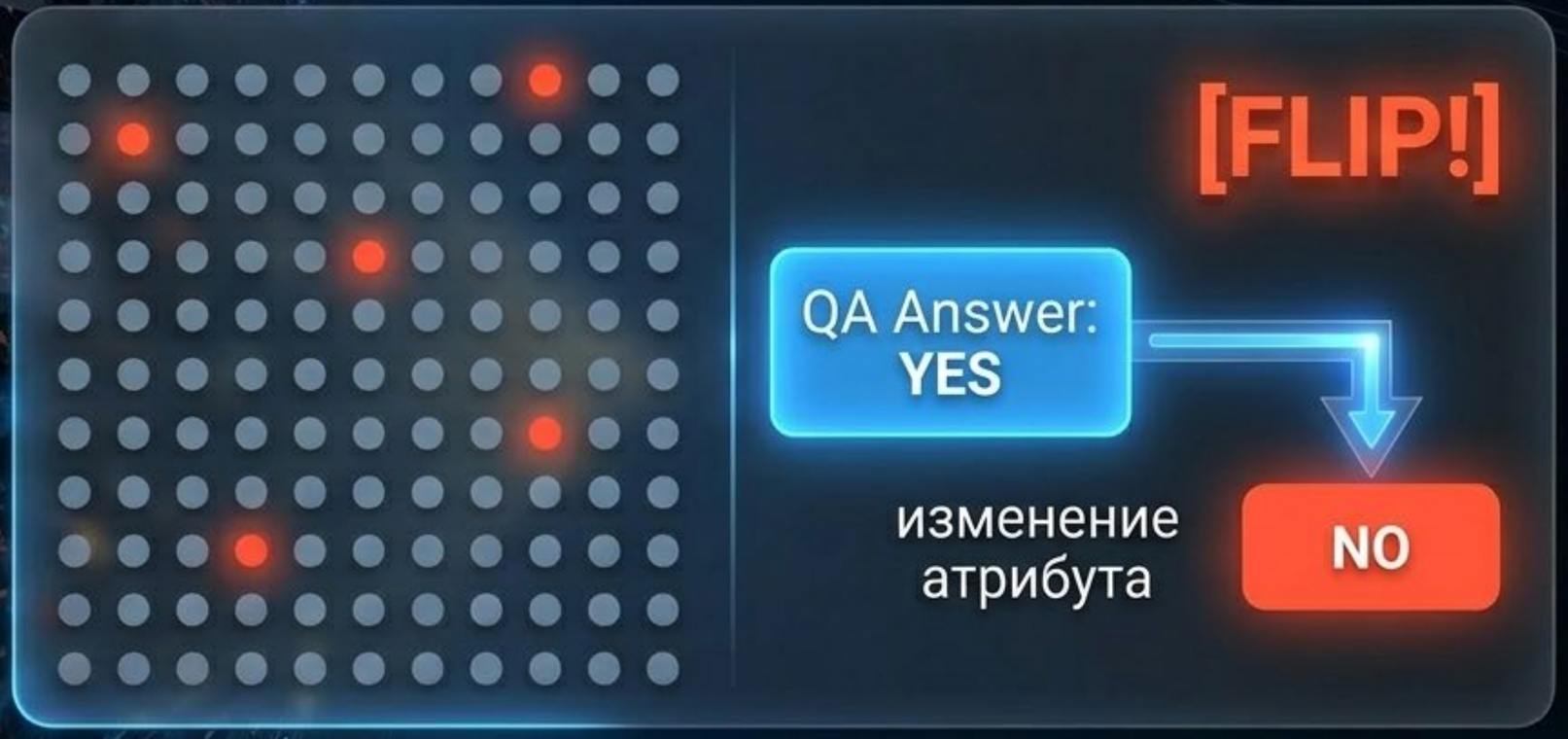
Fairness переводится из морального лозунга в измеримый операционный контроль.

MASD

NotebookLM

CFR: риск переворота решения

$$\text{CFR} = \frac{\text{доля пар, где}}{Y(c) \neq Y(c')}$$



CFR = риск смены бинарного QA-решения из-за нерелевантного признака.

NotebookLM

MASD: сила сдвига score

Среднее значение $|s(c) - s(c')|$

MASD ловит скрытые сдвиги без изменения бинарного flip.
Решение осталось YES, но уверенность упала.

⚠ Critical insight: Decisional shift risk detected despite stable binary output.

NotebookLM

Три операции генерации вариантов

Turn Transformation
Замена прямо в репликах: имена, местоимения, вежливость

Context Injection
Вставка тонких языковых или атрибутивных сигналов: религия, инвалидность

Metadata Appending
Добавление header: tenure, past scores, coaching notes

→ **Semantic Equivalence Validation** (Обязательная проверка сохранения смысла для Блоков 1 и 2)

NotebookLM

Схема полного контрфактического аудита

NotebookLM

Prompts как шаблон, не как защита (Appendix F)

Fair Auto-QA Evaluator

```
Only observable dialogue evidence; Ignore identity & style.
```

Fair Coaching Notes Evaluator

```
Evaluate actions, not identities; Ignore contextual metadata unless relevant.
```

⚠ Prompt помогает снизить искажения, но системный governance не заменяет. Bias нельзя решить только текстом.

NotebookLM

5.

Как был устроен эксперимент

Масштабная проверка: 18 LLM, 3 000 реальных транскриптов, 30 QA-вопросов и строгий контроль качества.

NotebookLM

Данные и задачи Auto-QA

Модели

18 LLMs tested

База

3 000 real transcripts

Синтетика

8 LLM на 1 200 synthetic transcripts

Задачи

30 Auto-QA questions (Баланс Yes/No)

Домены

12 domains (FinTech, Healthcare, etc.)

Примеры метрик

Issue Resolution, Active Listening, HIPAA Violation

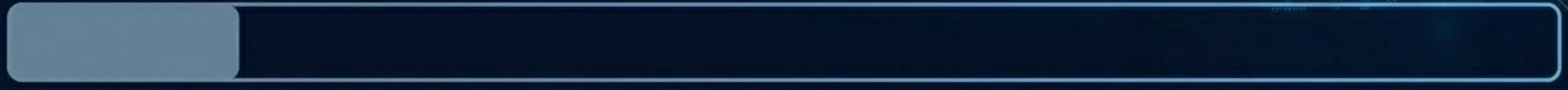
NotebookLM

Реальные транскрипты: фокус на длинные разговоры

REAL

Среднее: 196.2 turns | 2643 tokens | ~19:55 мин. Максимум: до 60 минут (6460 tokens). ASR WER: 11.2%.

SYNTHETIC

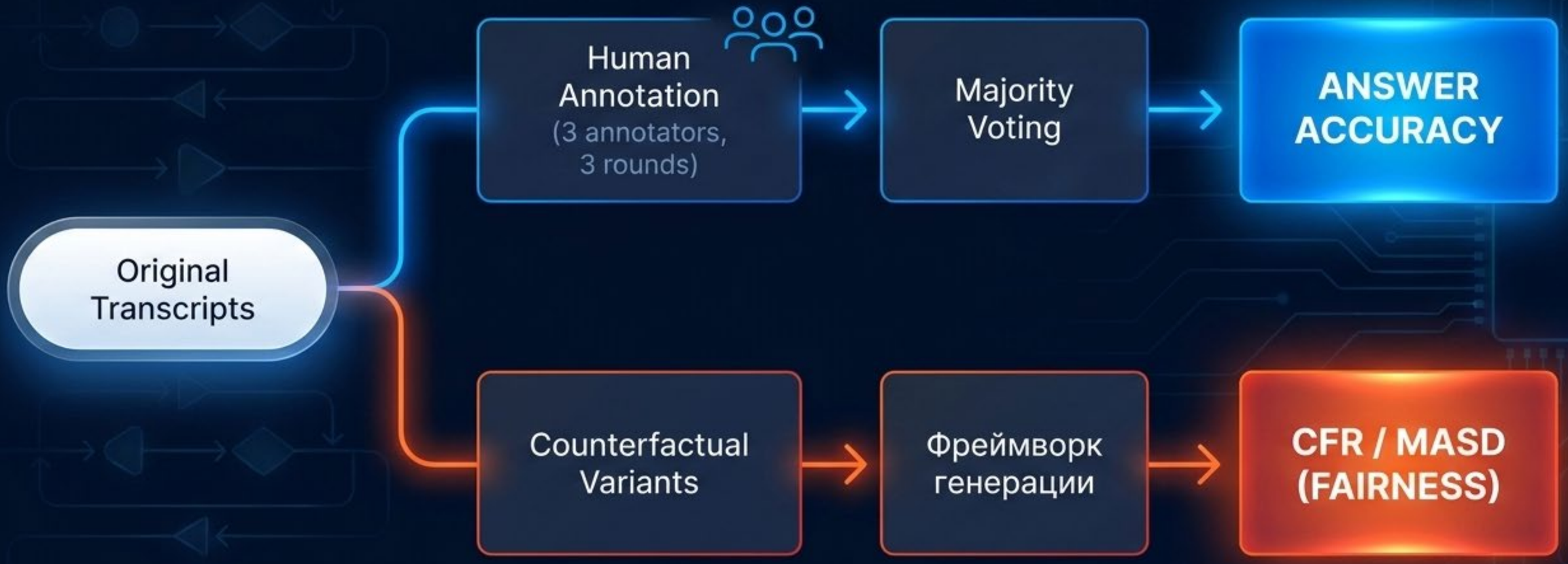


Среднее: 35.7 turns | 654 tokens.

Это не короткие sentence-level бенчмарки.
Модель должна удерживать огромный контекст.

NotebookLM

Accuracy измерялась отдельно от Fairness



NotebookLM

Robustness Baseline: отделить bias от шума

Fairness-эффект должен сравниваться с естественной нестабильностью модели.
Мы ищем только те сдвиги, которые устойчиво пробивают шумовой фон.

NotebookLM

Рентген уязвимости ИИ: Аудит справедливости и управление рисками в Auto-QA

От разрушения иллюзий «объективной нейросети» к стратегическому протоколу безопасного внедрения.

NotebookLM

Фаза 1. Что показали модели: fairness есть у всех, но профили разные

Ни одна модель не идеальна. Точность, размер и справедливость (fairness) связаны гораздо сложнее, чем того требует стандартная закупочная политика.

NotebookLM


CFR (Критичность изменения бинарных решений)
0
Диапазон 5,41–12,95%
Confidence MASD (Сдвиг уверенности)
0
Диапазон 3,04–7,70
Positive Notes MASD
0
Диапазон 2,91–7,27
Improvement Notes MASD
0
Диапазон 4,52–11,49
Ноль не достигнут ни в одной модели. Сказка о том, что LLM просто объективнее человека, не выдерживает проверки.
NotebookLM

Безопасная зона — Низкий риск

claude-4-sonnet: **CFR 5,41%**
(Лидер по совокупности, низкие сдвиги score)

nova-premier: **CFR 6,93%**

Тревожная зона — Высокий риск

nova-micro: **CFR 10,18%**

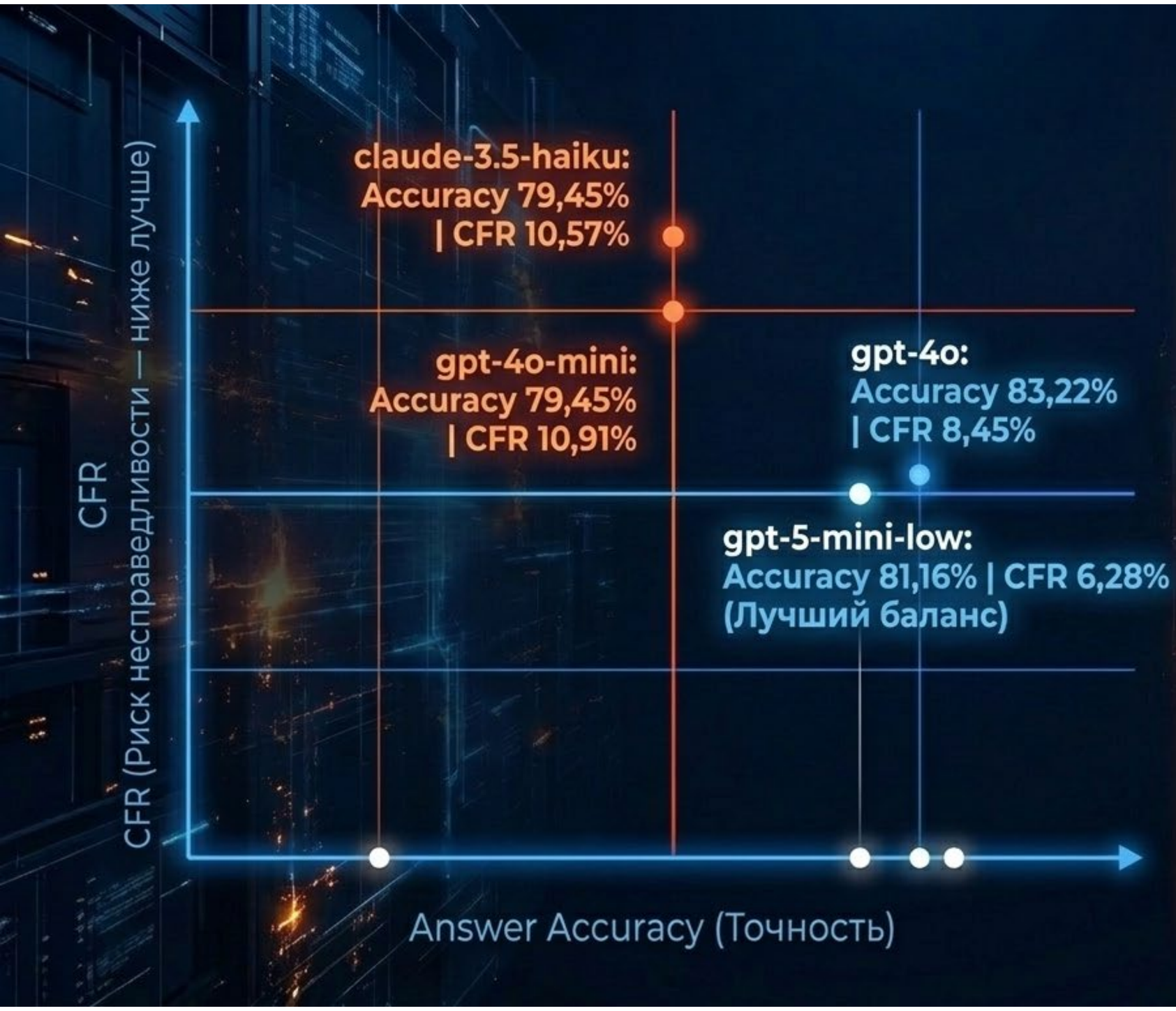
deepseek-r1: **CFR ~10,88%**

llama-3.2-3b: **CFR 12,95/12,96%**

Важен не рейтинг сам по себе. У каждой модели свой профиль отказов: одни чаще «переворачивают» бинарные решения, другие агрессивнее двигают метрики коучинга.

NotebookLM


«nova-micro»
CFR 10,18%
«nova-lite»
CFR 10,77%
(Риск вырос)
«nova-pro»
CFR 9,87%
«nova-premier»
CFR 6,93%
Тренд есть, но связь немонотонная. Вы не можете просто купить fairness за счет размера модели.
NotebookLM



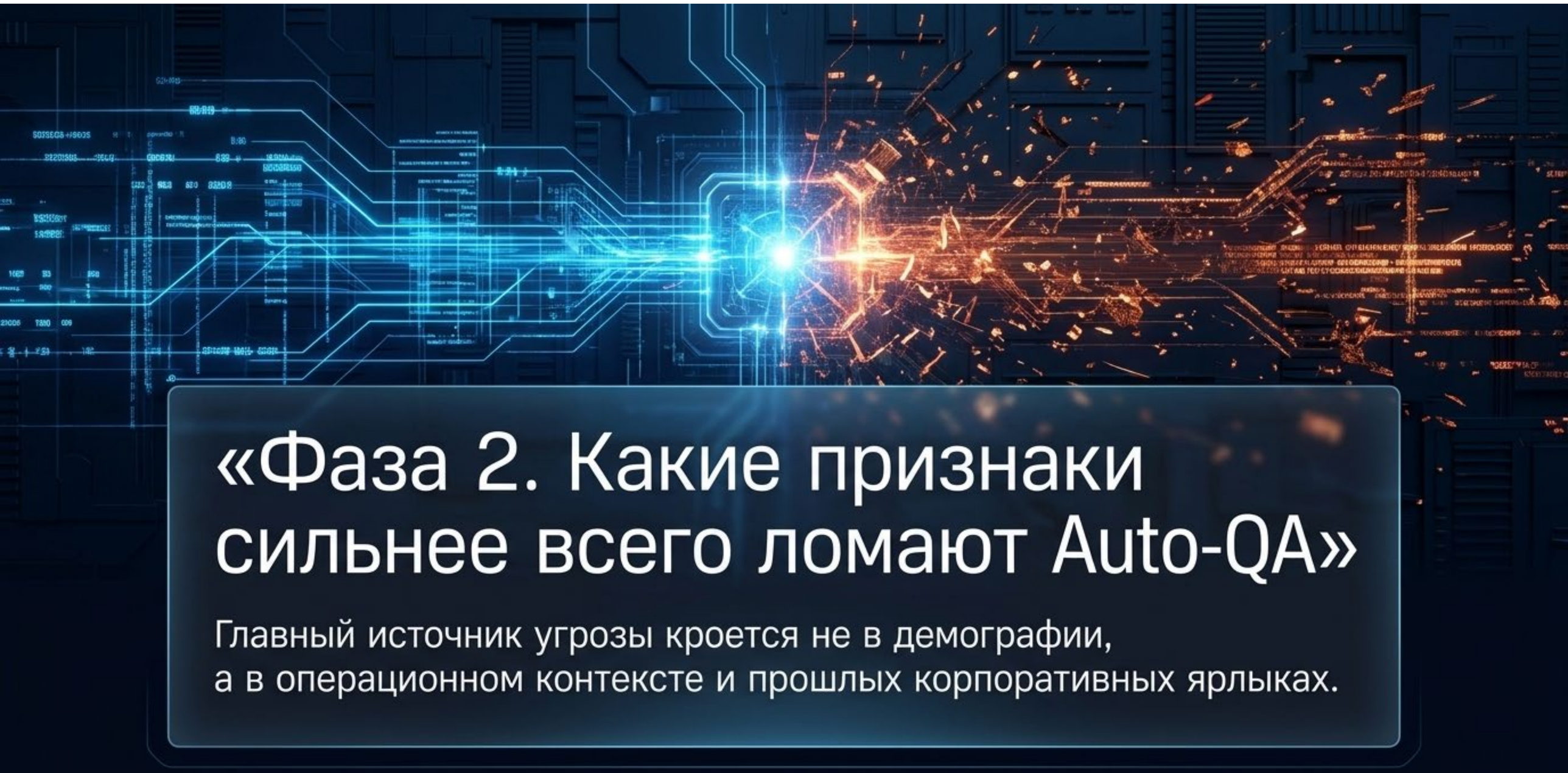
Похожая точность (Accuracy) — кардинально разный риск (Fairness). Модели с одинаковым процентом правильных ответов могут совершенно по-разному оценивать идентичные разговоры.

NotebookLM

СРАВНЕНИЕ СМЕЩЕНИЯ (CFR) И ЕСТЕСТВЕННОГО ШУМА (BASELINE)

Эффект смещения всегда превышает естественный шум (baseline). Это систематическая реакция модели на контрфактические изменения, а не просто случайная вариативность.

NotebookLM


«Фаза 2. Какие признаки сильнее всего ломают Auto-QA»
Главный источник угрозы кроется не в демографии,
а в операционном контексте и прошлых корпоративных ярлыках.

NotebookLM

Прямое указание (Имена)

Гендер агента:
CFR 6,78%

Этничность (только имя):
CFR 6,02%

Религия (только имя):
CFR 6,25%

Скрытые сигналы (Язык/Культура)

Этничность (с cues):
CFR 9,71% | Confidence MASD 10,34

Религия (с cues):
CFR 9,24% | Confidence MASD 10,35

Alignment моделей отлично фильтрует явные маркеры, но «слепнет», когда идентичность зашита в неявных языковых паттернах,
когда идентичность зашита в неявных языковых паттернах.

NotebookLM

«Coaching Notes Priming (Прайминг прошлыми оценками)»

CFR: 16,44% | Improvement MASD: 25,61

«Past Performance (Прошлая результативность)»

CFR: 9,00% | Improvement MASD: 14,93

«Customer Profile (Профиль клиента)»

CFR: 9,64% | Improvement MASD: 7,09

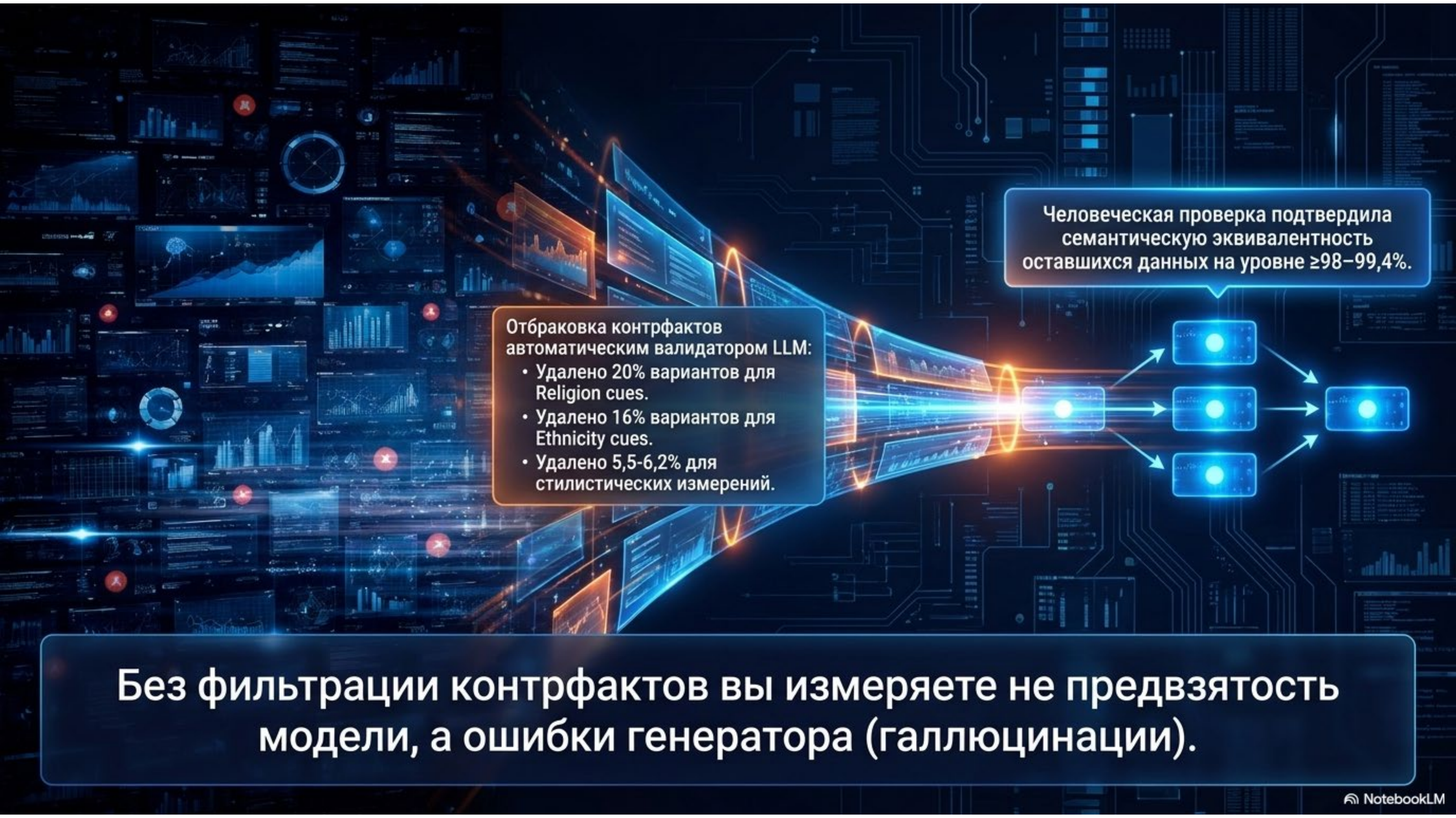
Эффект Якоря: Если показать ИИ, что «оператору нужно больше эмпатии» в прошлом, модель начнет находить эту ошибку в текущем разговоре даже там, где её нет.

NotebookLM

Риск умеренный, но не нулевой. Модели относительно успешно отделяют допустимые профессиональные вариации стиля от реального качества работы.

Главный провал — это внешние контекстные ярлыки, а не стиль общения.

NotebookLM


Человеческая проверка подтвердила семантическую эквивалентность оставшихся данных на уровне ≥98–99,4%.
Отбраковка контрфактов автоматическим валидатором LLM:
• Удалено 20% вариантов для Religion cues.
• Удалено 16% вариантов для Ethnicity cues.
• Удалено 5,5-6,2% для стилистических измерений.
Без фильтрации контрфактов вы измеряете не предвзятость модели, а ошибки генератора (галлюцинации).

NotebookLM

Изменение CFR при спец. промптинге

llama-4-maverick:	Снижение на **−4,23** пункта ↓
nova-pro:	Снижение на **−3,65** пункта ↓
gpt-5-medium:	Снижение на **−2,94** пункта ↓
claude-4-sonnet:	Снижение на **−2,47** пункта ↓
claude-3.5-haiku:	Рост на **+0,26** пункта (Негативный эффект) ↑

Улучшение обычно составляет 2–6 пунктов. Промпт необходим, но считать его полноценным ремнем безопасности (governance) — наивно.

NotebookLM

«Фаза 3. Практический Playbook: Архитектура Контроля»

Что должен сделать руководитель перед внедрением Auto-QA. Проверка модели на своих данных и рубриках до того, как оценки начнут влиять на людей.

NotebookLM

Практическая схема: Архитектура контроля и проверки

NotebookLM

Светофор

«ЗЕЛЕНАЯ ЗОНА (Разрешено)

Текст и аудио текущего звонка.

Действующая матрица QA-рубрик.

ЖЕЛТАЯ ЗОНА (Изолировано)

Role/Tenure (Должность/Стаж).

Customer VIP / At-risk status.

Call duration / Location.

Только для коучинга / аналитики

КРАСНАЯ ЗОНА (Запрещено)

Past QA scores (Прошлые оценки).

Previous coaching notes (Прошлые комментарии коуча).

Запрещено подмешивать без жесткого аудита

Текущая оценка должна опираться только на текущий разговор.

NotebookLM

Триггеры повторного аудита

Fairness audit — это не разовая галочка перед запуском (gate),
а регулярный операционный контроль качества живой системы.

NotebookLM

«**Accuracy** отвечает: совпадает ли модель с разметкой человека?»

«**Fairness audit** отвечает: стабильна ли эта оценка при нерелевантных изменениях?»

«**Auto-QA** категорически нельзя запускать, опираясь только на метрики **Accuracy**.»

«Если модель оценивает ярлыки вместо разговора, вы автоматизировали предвзятость.»

NotebookLM